DOCUMENT RESUME

ED 430 040 TM 029 784

AUTHOR Bertrand, Richard
TITLE IRT Design for the School Achievement Indicators Program (SAIP).
PUB DATE 1999-00-00
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Achievement Tests; Elementary Secondary Education; Foreign Countries; Item Bias; *Item Response Theory; *Mathematics Tests; Models; Test Items
IDENTIFIERS Canada; Dimensionality (Tests); Educational Indicators; *School Achievement Indicators Program (Canada); *Two Stage Testing

ABSTRACT

New ways are proposed to look at data from the Canadian School Achievement Indicators Program (SAIP) using item response theory (IRT) modeling. The focus is on the traditional test of the SAIP 1997 mathematics study. The test is two-staged in that the first 15 items, of median difficulty, were to be completed beforehand by all students as a "placement" test. Where students were to start the second portion of the test was determined by their scores on the first 15 items. The dimensionality of the scale, the item fit, item bias, and item invariance were considered in applying IRT modeling. The scale obtained reflected the relative ability of the students. Results suggest that mean scale scores can be safely compared between jurisdictions, genders, or age groups. The proposed approach seems reasonably appropriate when a two-stage testing procedure has been used to evaluate students' performance. (Contains 1 table, 5 figures, and 19 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 040

IRT Design for the School Achievement Indicators Program (SAIP)

Richard Bertrand

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
R. Bertrand
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM029784

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in Montreal Canada
April 1999

BEST COPY AVAILABLE

IRT design for the School Achievement Indicators Program (SAIP)

Richard Bertrand, Laval University
Léo Laroche, Council of Ministers of Education, Canada

Introduction

It was about ten years ago that the Canadian Council of Ministers of Education approved the implementation of a national assessment program called SAIP, the School Achievement Indicators Program. The aim of the program was to undertake large scale studies in mathematics, reading and writing and science for the 13 and 16-year-old French or English students. Two cycles of studies were anticiped for these three subject matters before year 2000 : the first cycle included mathematics in 1993, reading and writing in 1994 and science in 1996 ; for the second cycle, mathematics in 1997, reading and writing in 1998 and science in 1999. The 1999 science assessment is well under way and the Council has announced that the next cycle will be math assessment for 2001, reading and writing for 2002 and science for 2003.

Up till now the SAIP results had always been reported using classical statistics : proportions of students reaching a specific performance level for each age group (13 and 16 year-old), each gender and each jurisdiction (in Canada there are 10 provinces and three territories (with the new Nunavut)). The main objective of this paper is to propose new ways to look at the SAIP data using item response modelling.

SAIP features for the 1997 math assessment

SAIP follows two main goals the first of which is to evaluate the level of performance for Canadian students in the three subject matters. The second goal is to evaluate, between cycles of studies, changes in performances for each subject matter, that is to evaluate the so-called trend indicators.

The *usual* SAIP sampling design involves, for each jurisdiction, samples of about 500 students of each gender, each age level (13 or 16 year old) and each language (French or English). For example, the 1997 math study used a sample of about 45 000 students. A two-stage stratified cluster design was used, the first stage of sampling being the schools, proportional to the number of the students in the schools. The students were chosen at the second stage of sampling.

Each SAIP study involves two types of achievement variables : in mathematics, there is a traditional test made up of multiple choice and short answer items and a problem-solving test made up of just a few extended-response questions. For a usual SAIP study, the sample of students is divided in two subsamples, about half of the whole sample of students assigned to each type of achievement variables involved.

This paper will be dealing mainly with the traditional test of the 1997 mathematics study. This test was made up of 125 items, 60% of which were multiple-choice. There were 32 algebra items, 37 geometry items, 39 number concepts items and 17 statistics items. Each item was linked to one of five levels of performance defined beforehand, so that level 1 items were the easier ones and level 5 items the harder ones. Level 1 items were those from 16 through 40, level 2 items from 41 to 65, level 3 items from 1 to 15 and 66 to 75, level 4 items from 76 to 100 and level 5 items from 101 to 125.

The two-stage testing design

The 1997 math test can be considered two-staged in that the first 15 items, of median difficulty, were to be completed beforehand by all students. These 15 items formed what is called a placement test (Figure 1). Those having succeeded in not more than 10 of those 15 items were to carry on beginning with item 16 (the first encountered level 1 item) up to, if possible, item 125 (the last one); those succeeding in 11, 12 or 13 items of the first 15 items had to go to item 41 (the first encountered level 2 item) up to, if possible, item 125 and those having got more than 13 items right started at item 66 (the first encountered level 3 item, excluding items 1 to 15) up to, if possible, item 125. This two-stage feature created three non equivalent groups of students : group 1 being the less able students and group 3 the more able. While the majority of the 13-year-old students were in group 1, the majority of the 16-year-old students went in group 3.

[Insert Figure 1 about here]

In finding a way to scale the students and the items, we had to consider the fact that the three groups of students did'nt really go through the same test. We ended up using a three-parameter IRT non-equivalent group model as implemented in BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Apart from the information obtained from the items (difficulty, discrimination, pseudo-guessing), this model uses the information that the students are associated with one of three groups defined above, that is what Mislevy (1987) called auxiliary information, to estimate the ability of the students.

We argue that each student took *in fact* a 75-item test : level 1, level 2 and level 3 items for group 1 (G1) students ; level 2, level 3 and level 4 items for group 2 (G2) students ; level 3, level 4 and level 5 items for group 3 (G3) students.
Why is that so? The instructions to the students after the first 15 items of the placement test were :
for G1 (less able) students; start from item 16 (level 1 item) and hopefully go to item 125.
for G2 (median ability group) students; start from item 41 (level 2 item) and hopefully go to item 125.
for G3 (more able) students; start from item 66 (level 3 item) and hopefully go to item 125.

So that G3 students, the more able, were really administered a 75-item test. Now it is very doubtful that less able students (ex. G1 students) could do *conscientiously* more items than the more able G3 students in the same period of time (that is 2 hours and a half). In fact, we observed a whole bunch of missing values (sometimes more than 50%) corresponding to level 4 items and level 5 items for G1 students and also corresponding to level 5 items for G2 students. Besides, for the non-missing level 4 and level 5 items, the open-ended items were dramatically less successfull than the multiple choice items for G1 students; also, for the non-missing level 5 items, the open-ended items were less successfull than the multiple choice items for G2 students. Finally, a look at Sato's modified caution index[1] (McArthur, 1987) showed that the response patterns for G1 and G2 students answering the last 45 multiple choice items (that is level 3, 4 & 5 multiple choice items) were found more atypical (Figure 2) than the response patterns for G3 students, an evidence that many of these items could have been answered at random by G1 and G2 students.

[Insert Figure 2 about here]

Now these 75 items resulted in fact from the two tests of the two-stage design : the placement test made of 15 multiple choice items of median difficulty and the evaluation test (3 three forms) made up of the other items : 60 items in each form, one form for each group of students. The IRT model did fit much better and the unidimensionality was much easy to assume using only the items of the evaluation test : it seems that the items of the placement test would add some other dimension to the data[2]. The result of this was to actually retain only the items of the evaluation test to develop the scale.

Looking at the assumptions

We will be dealing with the dimensionality of the scale, the item fit, the item bias issue and add some considerations on item invariance.

To analyse the dimensionality of the IRT scale, we used full information item factor analysis (Bock, Gibbons & Muraki, 1988) as implemented in TESTFACT (Wilson, Wood & Gibbons, 1991). First, we performed a TESTFACT analysis by using all three groups of students and all items of the evaluation test. We also did a TESTFACT analysis for each of the three groups and their associated 60 items. In each occasion we ended up with two statistically significant factors[3] ; but

---

[1] This modified caution index was proposed by Harnish & Linn (1982). This index measures, for each response pattern, the departure from a perfect Guttman scale. When the items are ranked from the easiest to the hardest, a perfect pattern would be something like 1111000 : then the index gets a value of 0. Any departure from this perfect pattern causes the index to increase up to a maximum value of 1.

[2] DIMTEST was performed on each form of 75 items: the null hypothesis of one dominant factor was rejected for the 75 items of group 1. Moreover, as will be explained later, for the 15 items of the placement test, the IRT model did'nt fit the data.

[3] In a simulated study, De Champlain & Gessaroli (1998) found that TESTFACT revealed a large number of incorrect rejections of the assumption of unidimensionality. We used data sets already known as unidimensional (from IAEP study) and got also two factors based on item difficulty.

as the first factor was loaded with the easiest (or the hardest) items and the second factor with the hardest (or the easiest) items, we retained the 'one dominant factor' interpretation of Segall, Moreno & Hetter (1997). To validate that interpretation, we also looked at DIMTEST[4] statistics (Stout et al., 1992) for each of the three groups and their associated 60 items[5]. As can be seen from Table 1, each of these analyses revealed in fact only one dominant factor : two statistics were used for interpretation, Stout's T and Nandakumar's T; the null hypothesis of a dominant factor was not rejected (α=.05).

[Insert Table 1 about here]

As already been said, the item fit study (relying on visual inspection of the ICC's produced by BILOG-MG) was very much interesting in that it showed many items did'nt really fit the model when using together all items of the placement test and the evaluation test. For the 15 items of the placement test, the model was definitively not fitted to the data (Figure 3) while, for the other 110 items, the fit was generally better using only the evaluation test.

[Insert Figure 3 about here]

The DIF study was conducted looking first at the area between French and English item characteristic curves (Figure 4), that is the root mean square differences (RMSD) as suggested by Hulin, Drasgow & Parsons (1983, p.177). The inspection of the box plot of the RMSD's revealed two far outliers (Tukey, 1977). The same two items were found DIF using a procedure (Johnson, 1992) based on Mantel-Haenszel statistics. Following Camilli & Shepard (1994), these two DIF items were analysed by a panel of content experts, the result of which was to exclude these two items from the scale. The same two items were also found DIF using the data from the 1993 SAIP study[6].

[Insert Figure 4 about here]

To look at invariance of item parameters we first drew two random samples of students, each sample being made using students in each of the three non equivalent groups. Then, two other samples of students were drawn : the first one made up of the less able students and the second one made up of the more able students. Each sample involved more than 10 000 students. The items of the evaluation test were then calibrated for each of these samples. The correlation between the b values was very high using the two random samples (more than .9); but a modest correlation of .7 was found using the low ability and the high ability samples. Fan (1998) obtained much higher correlations (.87 and .96) between difficulty parameters in similar circumstances using low and high ability groups. We attributed this low correlation to the unexpected ceiling effect produced by the (too!) high ability group taking the more difficult items.

[Insert Figure 5 about here]

The IRT scaling

The preceding paragraphs discussed the assumptions underlying the construction of an IRT scale using the 108 non biased items of the evaluation test of the 1997 mathematics assessment. But is it worthwhile to report the SAIP results using such a scale? As already been said, SAIP traditionnaly reports student's results using 5 levels of performance : for example, to be considered at level 3 a student had to answer correctly to at least 15 items out of the 25 items of level 3 and to a minimum number of items in each strand (algebra, geometry, number concepts and statistics). Now the process of defining the levels of performance may be interesting but it is also very long, quite costly and probably a bit risky. IRT scaling does not involve this process of defining levels of performance. What if we used only classical number-right

---

[4] Mr Éric Frenette, research assistant at Laval University, performed the DIMTEST analyses.

[5] DIMTEST won't run with missing values and with more than 100 items; so using the three groups of students and all items produced a very large number of missing values: in fact, the number of students went from more than 23 000 to 4600; besides, of these 4600 non missing values, 4500 were G1 students. So we decided to rely only on the analyses involving each group of students and their associated 60 items. Because common items were linked to each group and that each DIMTEST analysis revealed a dominant factor, we felt that we could assume an *overall* dominant factor.

[6] But, at that time, no DIF study was done, so that these two items remained in the test for the 1997 SAIP study.

scores without bother with levels of performance? We argue that it would be quite reckless to use the classical number-right scores to compare student mean performances of the jurisdictions. In fact, we observed that the proportion of group 3 (more able) students varies a lot from one jurisdiction to another. But the ability of group 1 students having answered correctly 50% of the level 1, 2 and 3 items is by no means equal to the ability of group 3 students having answered correctly 50% of level 3, 4 and 5 items. The IRT modelling that was used here takes care of this nonequivalent ability feature between the students of the three groups. To show the huge gap between the classical score and the IRT scale score we looked at the frequency distribution of the students having succeeded in exactly 50% of the items, that is those having got a classical score of 50%. The IRT scale[7] scores for the 282 students having got the classical score of 50% vary from 309 to 611 : this is more than three standard deviation!

One of the drawbacks in using IRT scaling is of course to make much more difficult the comparison between the student's performance and the public's expectation[8]. On the other hand, scale anchoring methodology (King, Bertrand & Dupuis, 1989) could be used to help the interpretation of the results.

We are quite confident that the IRT scale so obtained reflects the relative ability of the students, even if they were assigned beforehand to one of the three non equivalent groups. We contend that mean scale scores could be safely compared between jurisdictions, genders or age groups. Two other features of this scale are worth mentionning. First, there were four strands of mathematics involved in the test : algebra, geometry, number concepts and statistics. The model used allowed for the development of four subscales, one for each of the strands. The subscales were built without recalibrating the items but used instead, as priors, item parameters already estimated. The subscales were then used to compare the mean scale scores of each jurisdiction : the latter comparisons being probably much appealing for the curriculum experts. Second, we wanted to look at trend indicators, that is to compare the 1993 results to the 1997 results. The scale was built using all students from the two cohorts and the unbiased items common to the 1993 test and the 1997 test.

Even if the nonequivalent group IRT scaling still raises some important and non answered questions, including item invariance and dimensionality issues, we argue that the approach presented here seems reasonably appropriate when a two-stage testing procedure has been used to evaluate student's performance[9].

[7] The mean of the scale is 500 and the standard deviation is 100.

[8] See for example the much debated issue of using achievement level setting in NAEP reports as discussed by Reckase (1998), Linn (1998) and Mislevy (1998).

[9] Author's e-mail: richard.bertrand@fse.ulaval.ca

References

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, **12**, 261-280.

Bock, R. D., & Zimowski, M. F. (1995). Multiple group IRT. In W. van der Linden & R. Hambleton (Eds.), *Handbook of item response theory*. New York : Springer-Verlag.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA : Sage.

De Champlain, A., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test tengths. *Applied Measurement in Education*, **11**, 231-253.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, **58**, 357-381.

Johnson, E. G. (1992). Theoretical justification of the omnibus measure of DIF. IAEP Technical report (Appendix 3). Princeton, NJ : Educational Testing Service.

King, B. J., Bertrand, R., & Dupuis, F. A. (1989). A world of differences : an international assessment of mathematics and science. Technical report. Princeton, NJ : Educational Testing Service.

Linn, R. L. (1998).Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education*, **11**, 23-47.

McArthur, D. L. (1987). Analysis of patterns : the S-P technique. In D. L. McArthur (Ed.), *Alternative approaches to the assessment of achievement*. Boston : Kluwer.

Mislevy, R. J. (1998).Implications of market-basket reporting for achievement-level setting. *Applied Measurement in Education*, **11**, 49-63.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, **11**, 81-91.

Reckase, M. D. (1998). Converting boundaries between National Assessment Governing Board performance categories to points on the National Assessment of Educational Progress score scale : The 1996 science NAEP process. *Applied Measurement in Education*, **11**, 9-21.

School Achievement Indicators Program. (1997). *SAIP Mathematics II*. Toronto, Council of Ministers of Education, Canada.

School Achievement Indicators Program. (1997). *SAIP Science : Technical Report*. Toronto, Council of Ministers of Education, Canada.

School Achievement Indicators Program. (1993). *SAIP Mathematics*. Toronto, Council of Ministers of Education, Canada.

Segall, D. O., Moreno, K. E., & Hetter, R. D. (1997). Item pool development and evaluaiton. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing : from inquiry to operation*. Washington, DC : American Psychological Association.

Stout, W. F., Nandakumar, R., Junker, B., Chang, H., & Steidinger, D. (1992). DIMTEST : A fortran program for assessing dimensionality of binary item responses. *Applied Psychological Measurement*, **16**, 236.

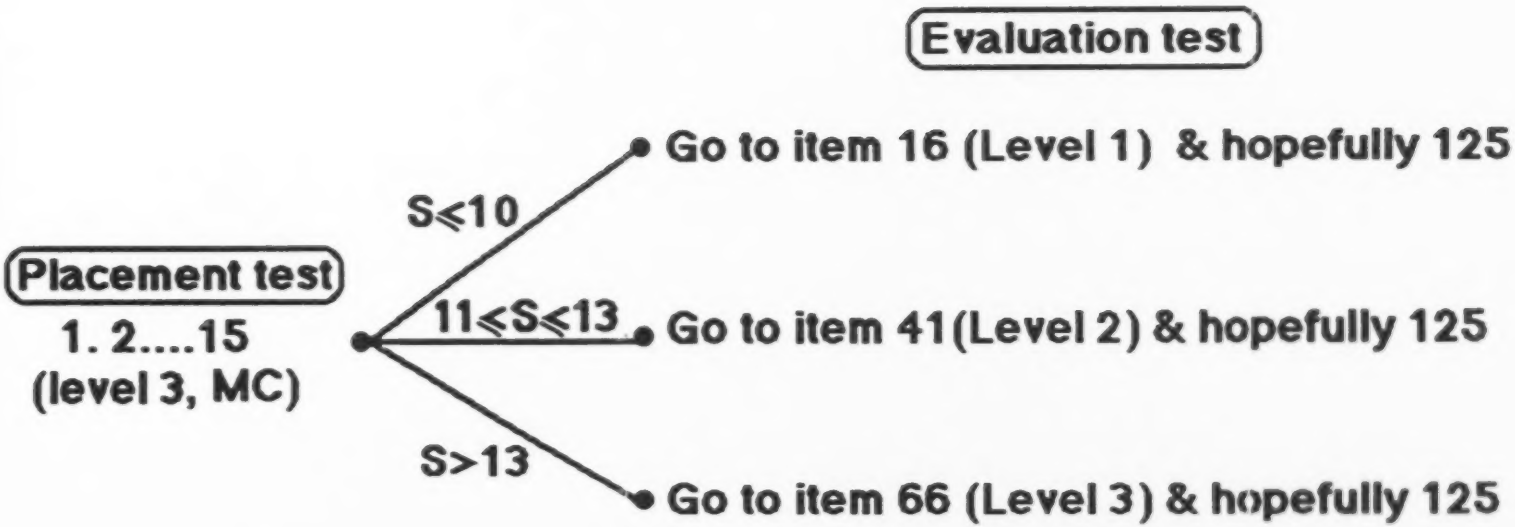
Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT. Test scoring, item statistics and item factor analysis.* Mooresville, IN : Scientific Software.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG. Multiple-group IRT analysis and test maintenance for binary items.* Mooresville, IN : Scientific Software.

Table 1
T statistics and probability p using Stout's and Nandakumar's method (DIMTEST)

Method	Group 1		Group 2		Group 3	
	T	p	T	p	T	p
Stout	-.832	.797	.614	.270	-.988	.838
Nandakumar	-.977	.836	.916	.180	-1.311	.904

BEST COPY AVAILABLE



Typically... for G1 students : item 1--- item 75
for G2 students : item 1--- item 15 & item 41 to item 100
for G3 students : item 1--- item 15 & item 66 to item 125

Figure 1 The two-stage testing procedure of the 1997 math assessment

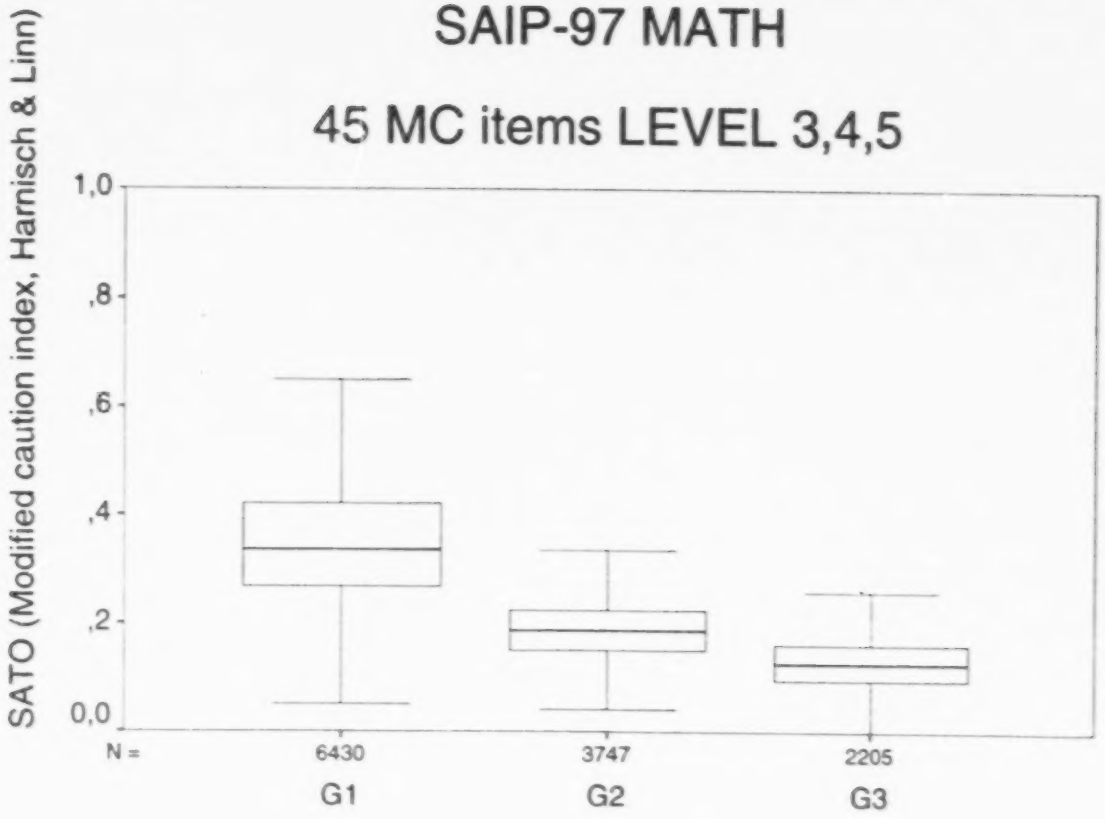
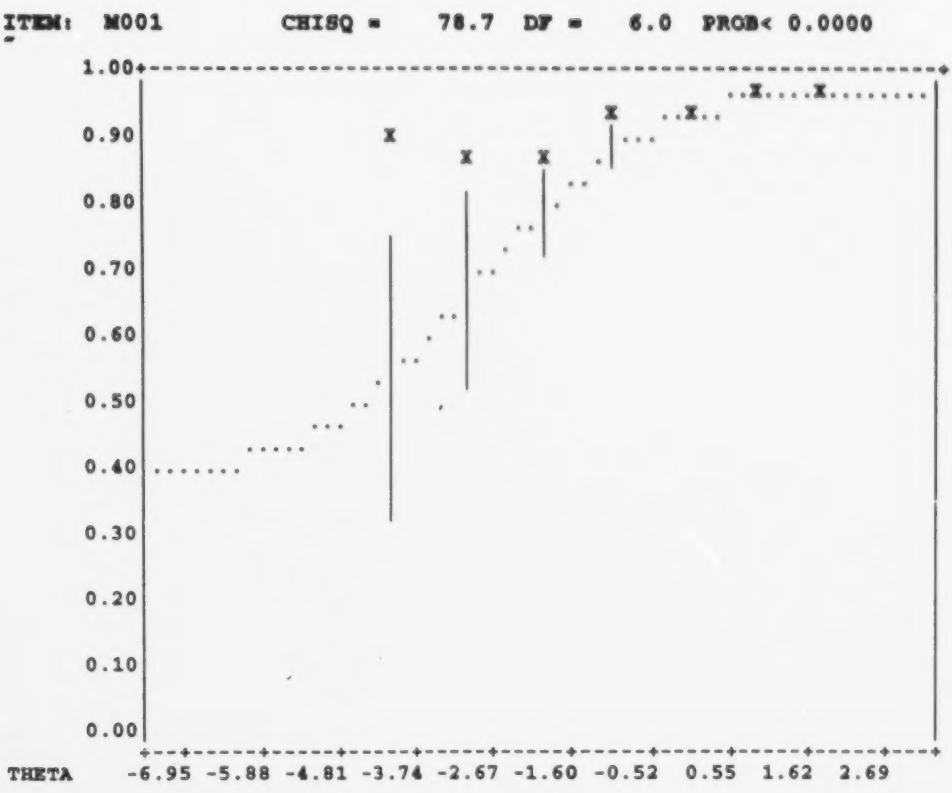


Figure 2 Sato's indices for three groups



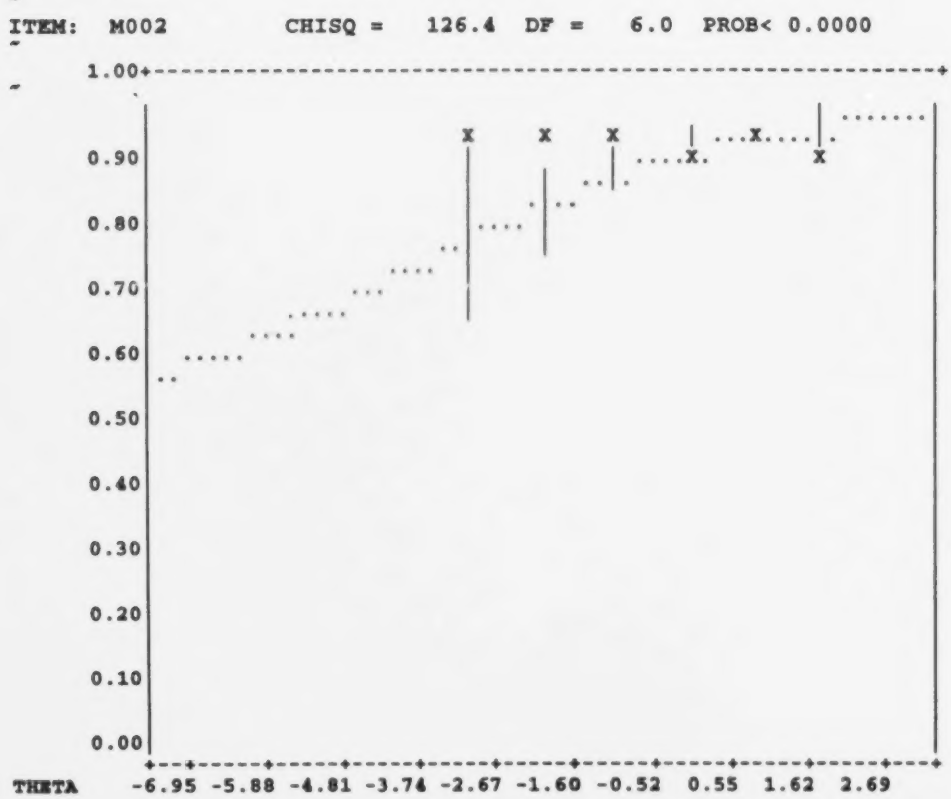
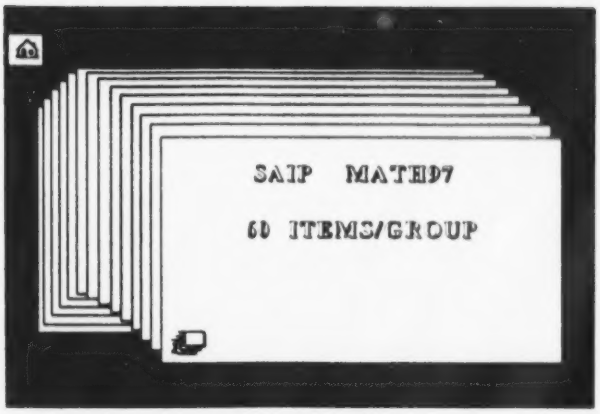
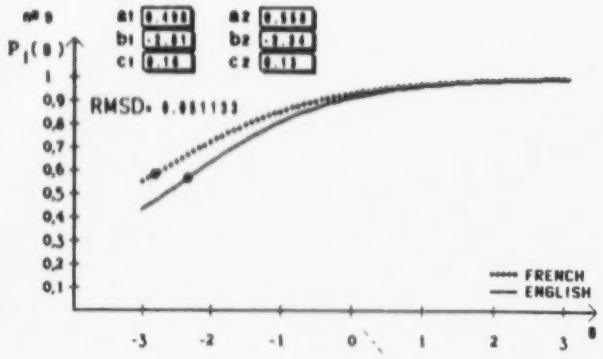
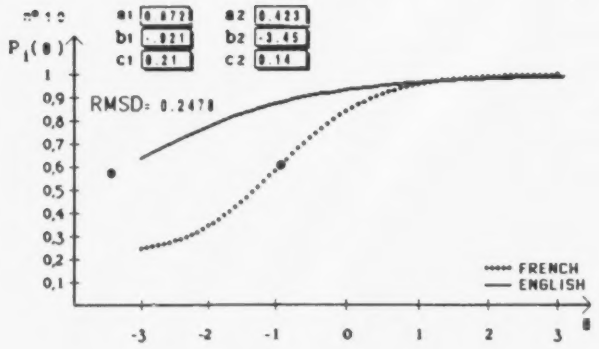


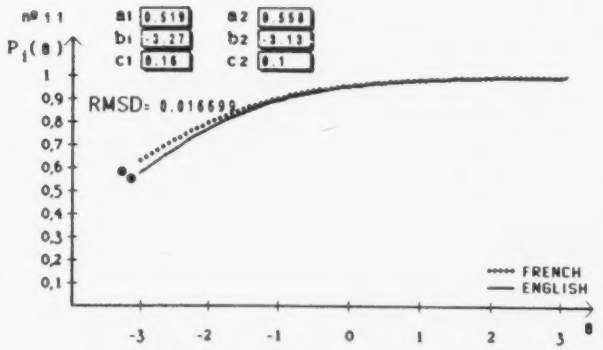
Figure 3 ICC's of Item 1 and Item 2 of the placement test (from BILOG-MG)

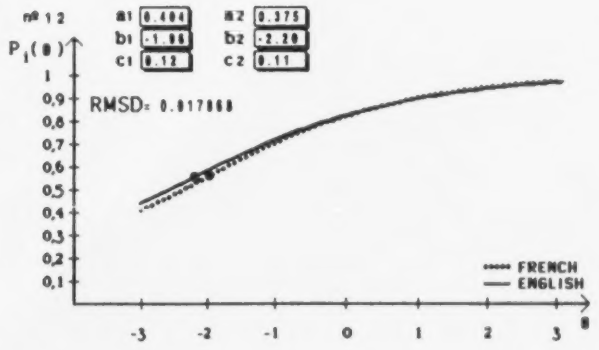
BEST COPY AVAILABLE

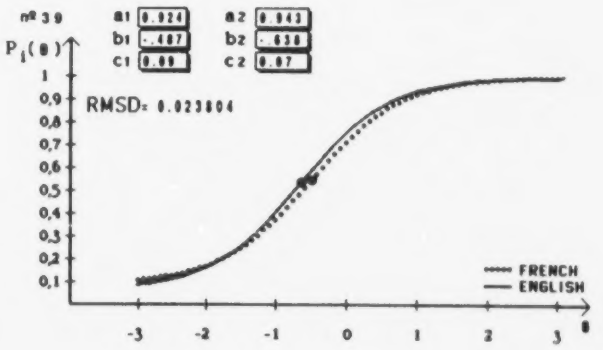


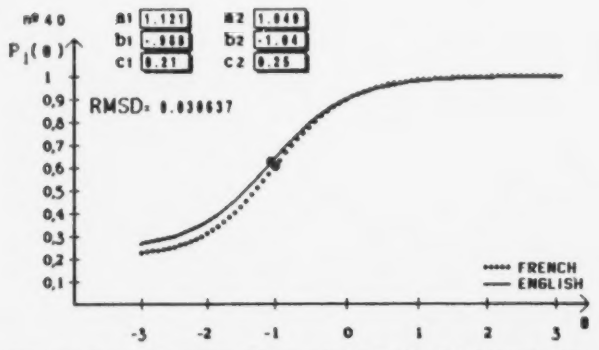












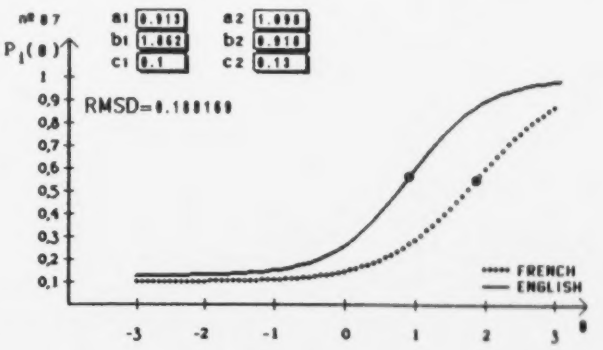
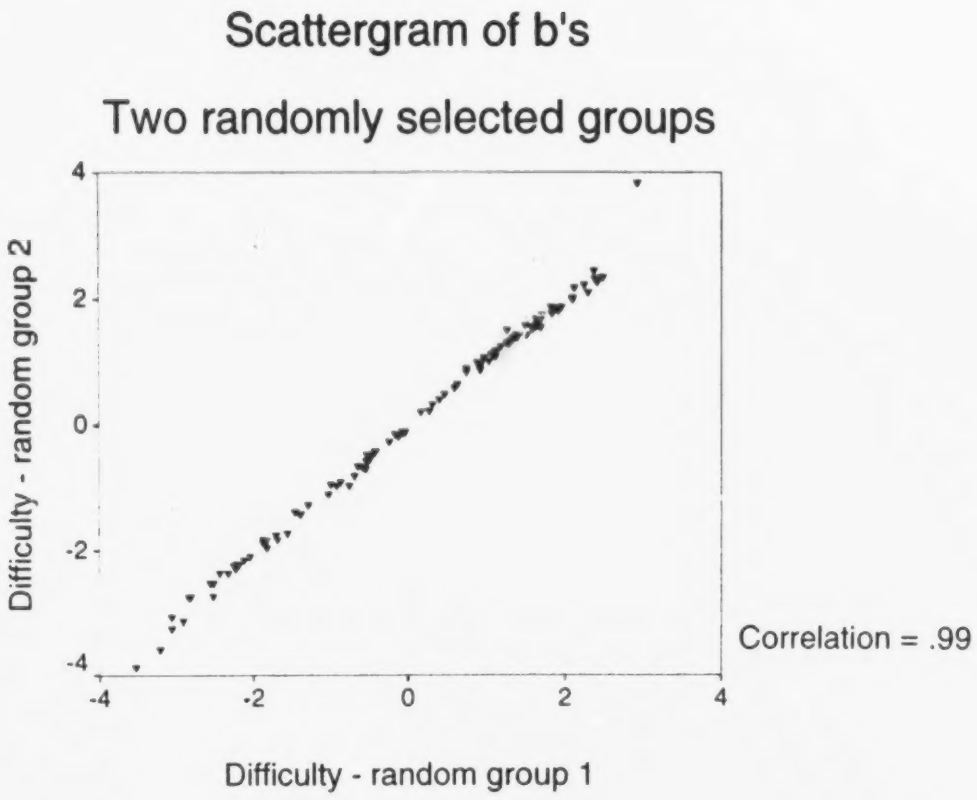


Figure 4 Root mean square difference between English & French ICC

BEST COPY AVAILABLE



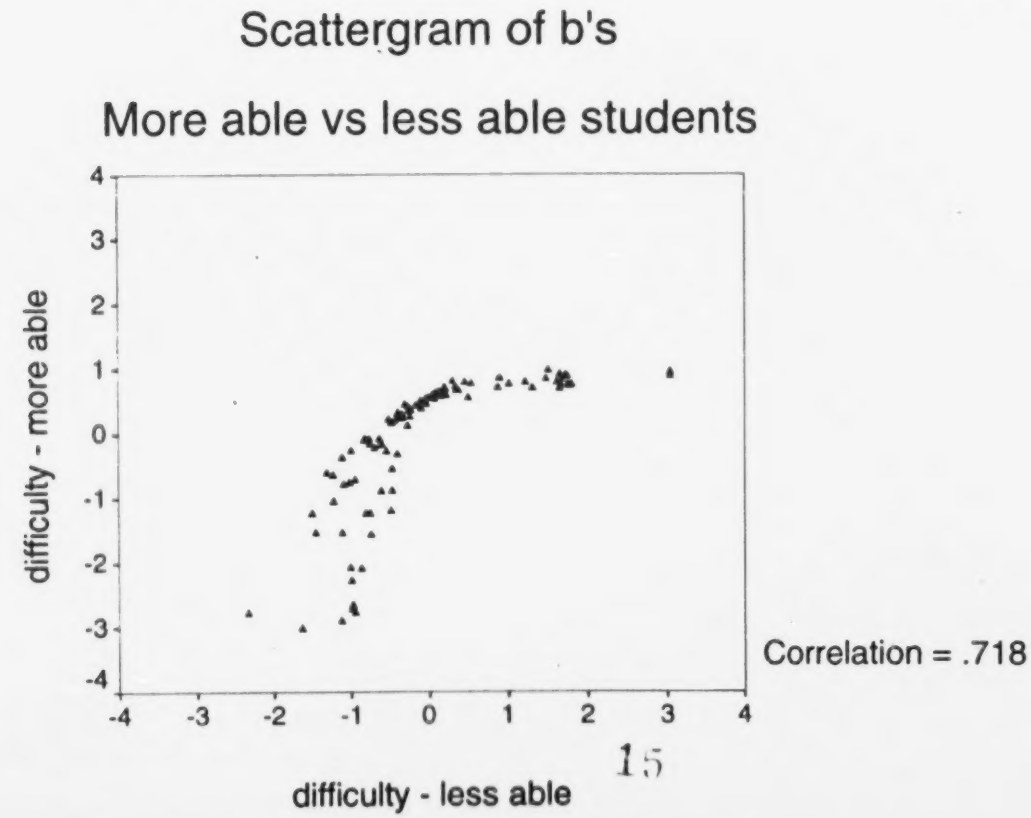


Figure 5 Scattergrams of difficulty indices (b) for two subgroups